

# Knowledge Discovery from Texts on Agriculture Domain

**Mathieu Roche**

**Cirad – TETIS and LIRMM – Montpellier, France**

**Web:** <http://www.textmining.biz>

**Email:** [mathieu.roche@cirad.fr](mailto:mathieu.roche@cirad.fr)



# Knowledge Discovery from Texts on Agriculture Domain



# Outline

- Part 1** Data Science and Big Data
- Part 2** Heterogeneity and textual data
- Part 3** Applications in agriculture domain
- Part 4** Conclusions and future work



Part 1

# Data Science and Big Data





# Big Data

**Volume**

**Velocity**

**Variety**

***3V of Big Data***

**Variability, Véracity, Value,**

**Visualisation, Valorization**





## Part 2

# Heterogeneity and textual data



```
0/1 x B_1404 [WARNING]: "Asynchronous reset/set/load <%item> exists in module/unit"
0/1 x B_1405 [WARNING]: "<%value> asynchronous resets in this unit detected"
0/1 x B_1406 [WARNING]: "<%value> synchronous resets in this unit detected"
0/1 x B_1407 [ERROR]: "Do not use active high asynchronous reset/set/load"
```

```
// Total Module Instance Coverage Summary
```

```
lines
statement
```

```
Policy:
<v
-----
0/1
```

PERCENT

31.54

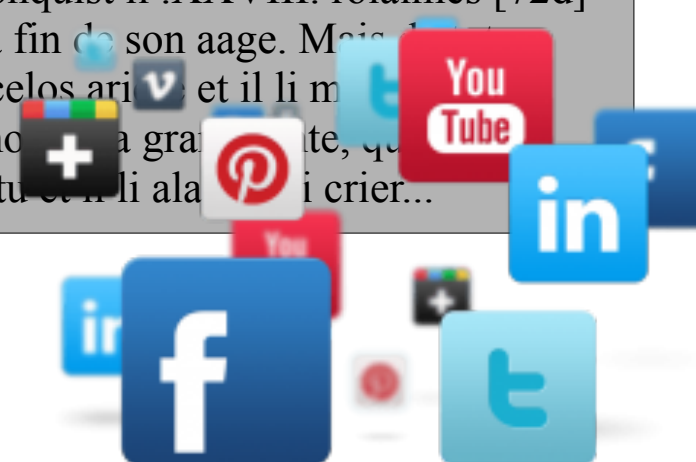
31.54

>]:<message>

is not allowed to be used as



descovri son corage a Lancelot et dist  
a guerre commença, baoit il a tot le  
e: et bien i parut, kar il fu a vint et cinc  
puis conquist il .XXVIII. roialmes [72d]  
ns fu la fin de son aage. Mais  
st Lancelos ari et il li m  
grant ho a la gra te, q  
e roi Artu et il li ala i crier...





## Textual data and satellite images

[Roche *et al.* SI'2014]

### Vakinankaratra – L'agriculture de conservation lancée

17.12.2014 | 7:18 | Non classé | 0



L'agro-écologie est une nécessité. Plus de 80% de la population malgache vit en milieu rural et opère en général dans l'agriculture. La croissance démographique associée au changement climatique provoque une forte destruction de l'environnement et une dégradation alarmante de la fertilité des sols. Afin d'y faire face et pour mieux lutter contre la malnutrition, le Groupement semis direct de Madagascar lance le projet Manitatra dans quatre communes rurales du district de Betafo et de Mandoto, dans la région Vakinankaratra. Ce projet est réalisé en partenariat avec le ministère de l'Agriculture et du développement rural et sur financement de l'Association française du développement et du Comesa.

Le groupement qui focalise son activité sur l'agro-écologie et l'agriculture de conservation, sensibilise et incite les paysans des communes ciblées à pratiquer l'agriculture sous couverture végétale et la rotation culturale. Et afin d'assurer une sécurité alimentaire de la commune rurale d'Ankazomiriotra, d'Inanantonana, de Vinany et de Fidirana, le projet Manitatra compte adhérer 1000 paysans, dont 200 femmes, sur la pratique de ce système de culture agro-écologique qui ne nécessite pas des nombreux travaux et éreintant comme l'exige le labourage. « Il suffit que les paysans recouvrent le sol de végétaux et cultivent sans dépenser du temps et de l'argent pour l'achat d'outils », ne Rakotondramanana, directeur exécutif du projet qui s'active aussi dans le Sud-Est de Madagascar. Des formations sur la régénération de la fertilité du sol et la lutte contre sa dégradation ainsi que l'introduction du système des légumineuses seront la priorité des activités du projet.

Angola Ny Avo



**Agritrop**

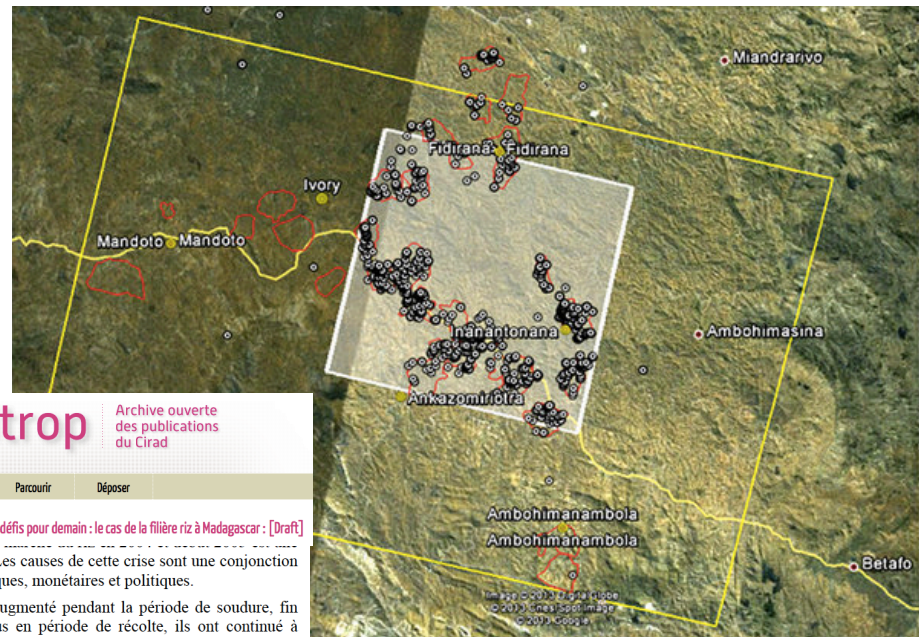
Archive ouverte des publications du Cirad



Crise hier, opportunités aujourd'hui, défis pour demain : le cas de la filière riz à Madagascar : [Draft]

augmentation sans précédent des prix de détail. Les causes de cette crise sont une conjonction de plusieurs facteurs, internes et externes : physiques, monétaires et politiques. L'an dernier, les prix du riz ont normalement augmenté pendant la période de soudure, fin 2003 début 2004, mais ne sont pas redescendus en période de récolte, ils ont continué à augmenter à un rythme soutenu. La variation annuelle du prix du paddy entre récolte et soudure est habituellement de l'ordre de 50% au Lac Alaotra, elle a été de 150% en 2004-2005 [Minten et Ralison, 2005]. Le prix du riz national ou importé dépassait historiquement 1000 Ar le kg entre septembre 2004 et février 2005 sur les marchés de la capitale. Si on compare l'évolution des prix du riz en 2001 et en 2004, on peut se rendre compte que les trois premiers mois de l'année il coûtait moins cher en 2004 qu'en 2001 et 2,5 fois plus cher en novembre.

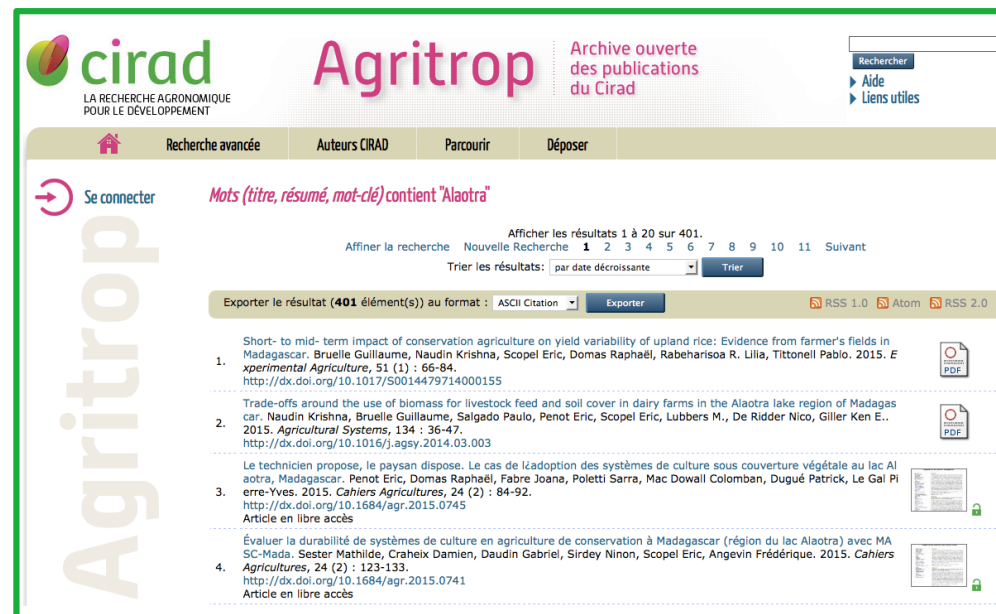
Cette hausse des prix s'est généralisée dans tout le pays. Elle s'est répercutée dans l'espace : marchés urbains et ruraux, auprès de tous les agents de la filière et pour toutes les variétés de riz (vary gazy, makalioka, tsipala, riz pluvial...). A titre d'exemple, dans le Moyen-Ouest, la hausse des prix du riz a été aussi importante sur les marchés situés en bord de route nationale que sur les marchés plus enclavés comme Inanantonana (45 mn de piste en saison sèche), Vasiana (1h15mn), Mahasolo (2h30mn) ou Ambalanirana (4h).



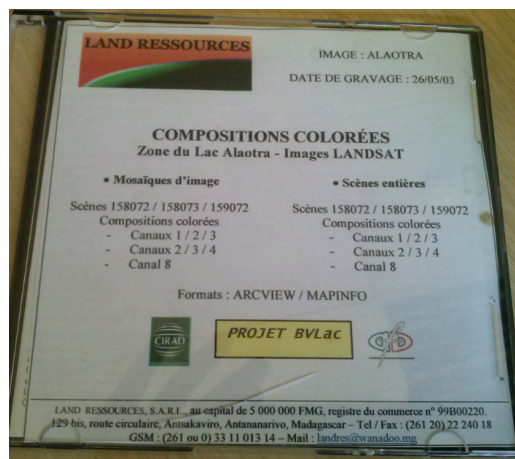
- Data and Issue



- Hard Disc (157 188 files)



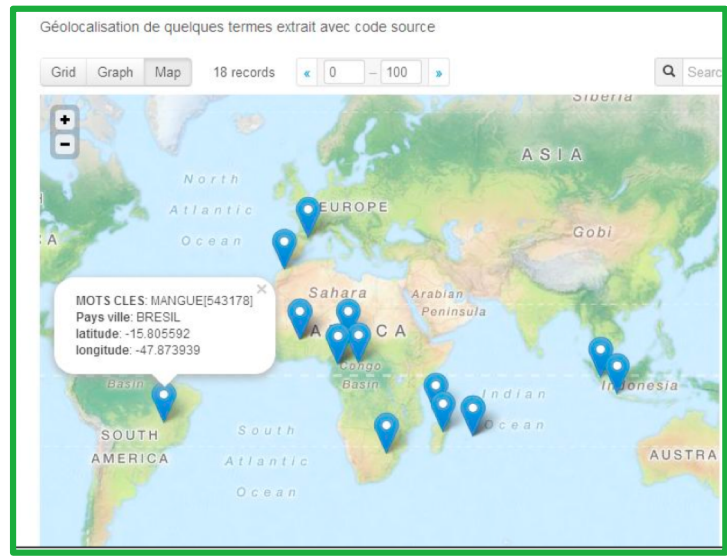
The screenshot shows the Agritrop website interface. At the top, there's a search bar and navigation links like 'Rechercher', 'Aide', and 'Liens utiles'. Below the header, there's a section for 'Se connecter' and a search filter 'Mots (titre, résumé, mot-clé) contient "Alaotra"'. The main content area displays a list of search results, including titles, authors, and publication details. For example, one result is 'Short- to mid- term impact of conservation agriculture on yield variability of upland rice: Evidence from farmer's fields in Madagascar' by Bruelle Guillaume, Scopel Eric, Domas Raphaël, Rabearisoa R. Lilla, Tittone Pablo. 2015. *Experimental Agriculture*, 51 (1) : 66-84. The results are sorted by date in descending order, and there are options to export the results in various formats like ASCII Citation, RSS 1.0, Atom, and RSS 2.0.



- Method: Extraction of features** [Roche *et al.* CA'2015]

## 3 types of features:

- thematic features
- spatial entities
- temporal entities



17.12.2014 | 7:18

Non classé

0




L'agro-écologie est une nécessité. Plus de 80% de la population malgache vit en milieu rural et opère en général dans l'agriculture. La croissance démographique associée au changement climatique provoque une forte destruction de l'environnement et une dégradation alarmante de la fertilité des sols. Afin d'y faire face et pour mieux lutter contre la malnutrition, le Groupement semis direct de Madagascar lance le projet Manitatra dans quatre communes rurales du district de Betafo et de Mandoto, dans la région Vakinankaratra. Ce projet est réalisé en partenariat avec le ministère de l'Agriculture et du développement rural et sur financement de l'Association française du développement et du Comesa.

Le groupement qui focalise son activité sur l'agro-écologie et l'agriculture de conservation, sensibilise et incite les paysans des communes ciblées à pratiquer l'agriculture sous couverture végétale et la rotation culturale. Et afin d'assurer une sécurité alimentaire dans la commune rurale d'Ankazomiriotra, d'Inanantonana, de Vinany et de Fidirana, le projet Manitatra compte adhérer 1000 paysans, dont 200 femmes, sur la pratique de ce système de culture agro-écologique qui ne nécessite pas des nombreux travaux et éreintant comme l'exige le labourage. « Il suffit que les paysans recouvrent le sol de végétaux et cultivent sans dépenser du temps et de l'argent pour l'achat d'outils », note Rakotondramanana, directeur exécutif du projet qui s'active aussi dans le Sud-Est de l'île. Des formations sur la régénération de la fertilité du sol et la lutte contre sa dégradation ainsi que l'introduction du système des légumineuses seront la priorité des activités du projet.





- (a) Extraction of features: **thematic terms** [Lossio Ventura *et al.* ISWC'2014]



**BioTex**

•Système de culture  
•Production  
•Développement durable  
•Eau ...

•Système de culture  
•Développement durable  
•Ressources naturelles  
•Mise en œuvre ...

Patterns Information

Number of linguistic patterns  [Patterns extracted from UMLS for English and Spanish, and from MeSH for French](#)  
*used to filter candidate terms*  
Ex: Noun Noun; Noun Prep:det Noun; ... [more examples](#)

Type of terms to extract

☐ All Terms ☒ Multi Terms  
*single-word + multi-word term* *multi-word term*

Measures selection and data


Select ranking measure  [read more](#)


Type of documents ☒ Single Document ☐ Set of Documents


File source  Aucun fichier sélectionné.  
*Only ".txt" accepted as file extension*


Language of your text

**Institutions**


 Laboratoire Informatique Robotique Microélectronique Montpellier

 UNIVERSITÉ MONTPELLIER

 cnrs

 TETIS

**Sponsors**

 SIFR project





- (a) Extraction of features: **spatial features (SF)**

## *Model*

- **Global Model:** SF is composed of at least one Named Entity (NE) and one variable number of spatial indicators specifying its location. SF can then be identified in two ways:
- **Absolute spatial feature (A\_SF)** one NE with a geo-localization, such as  $\langle (\text{spatialIndicator})^*, \text{NE of Location} \rangle$  (ex: *the city of Constantine*).
- **Relative spatial feature (R\_SF)** one spatial with at least one SF (ex: *in the south of the city of Constantine*).  
An R\_SF is defined as  $\langle (\text{spatial relation})^{1..*}, \text{A\_SF} \rangle$  or  $\langle (\text{spatial relation})^{1..*}, \text{R\_SF} \rangle$   
**Five spatial relation types are considered:** orientation, distance, adjacency, inclusion, and geometric which defines union or intersection linking two SFs.



- (a) Extraction of features: **spatial features (SF)**

**Methods** [Kergosien *et al.*, IJGIS'2014]

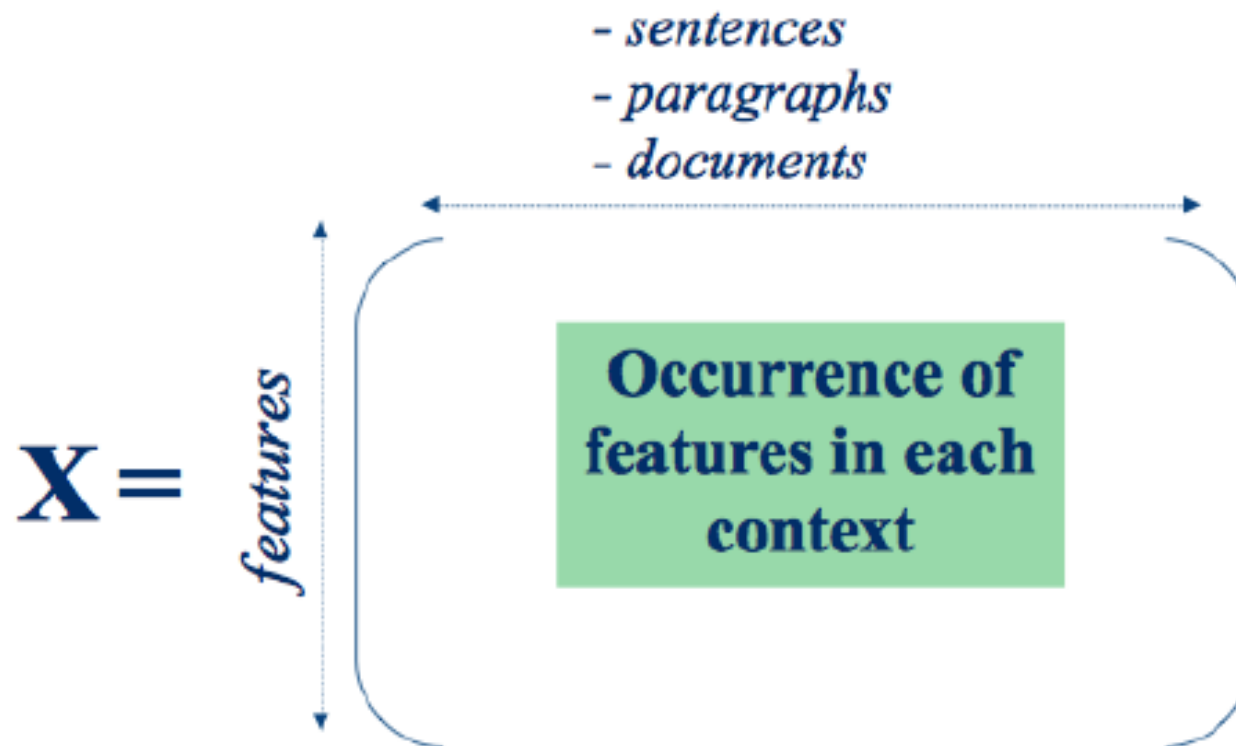
- **Symbolic approach**: Using rules (*Text2Geo*) for extracting A\_SF and R\_SF

Basic patterns			Text2Geo patterns			
	A_SF	R_SF		R_SF	R_SF	OE
Precision	20%	48%	Precision	53%	84%	92%
Recall	63%	27%	Recall	94%	66%	35%
F-mesure	30%	34%	F-mesure	67%	74%	50%

- **Statistic approach**: Using context and IR methods for spatial features disambiguation



- (b) Representation of documents



- (c) **Similarity**

$$\text{Global\_Sim}(\text{vect1}, \text{vect2}) = \alpha \cdot \text{cosT}(\text{vect1}, \text{vect2}) + (1-\alpha) \cdot \text{cosS}(\text{vect1}, \text{vect2})$$

with  $\alpha \in [0, 1]$

**cosT**: cosine based on **thematic features** (BioTex)

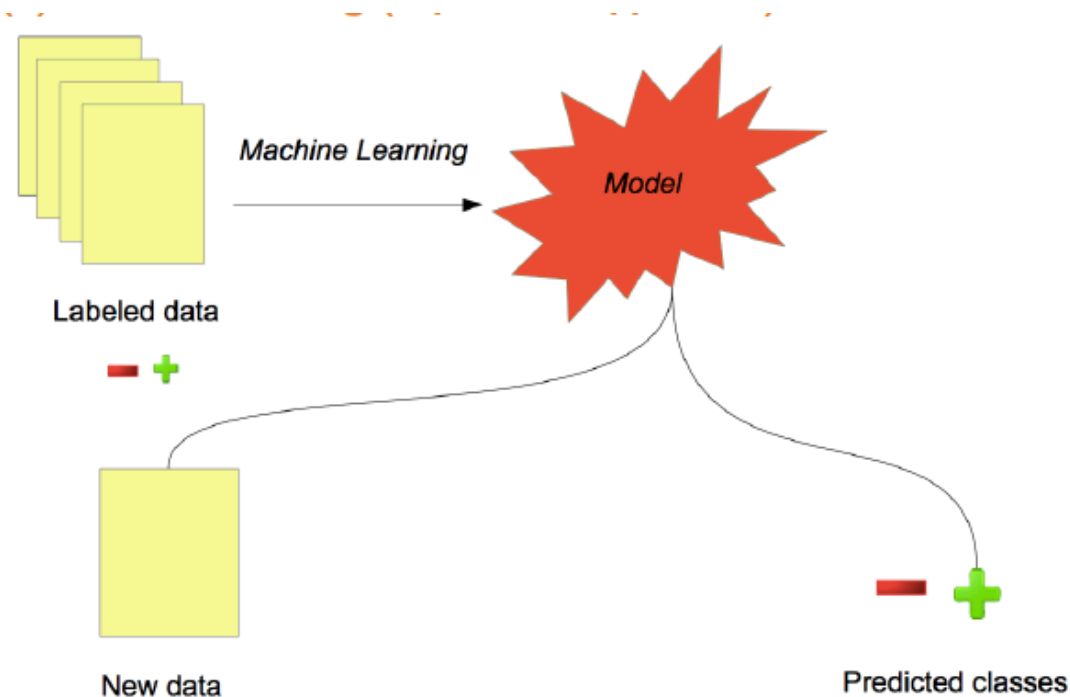
**cosS**: cosine based on **spatial features**

**Perspective:** adding temporal information

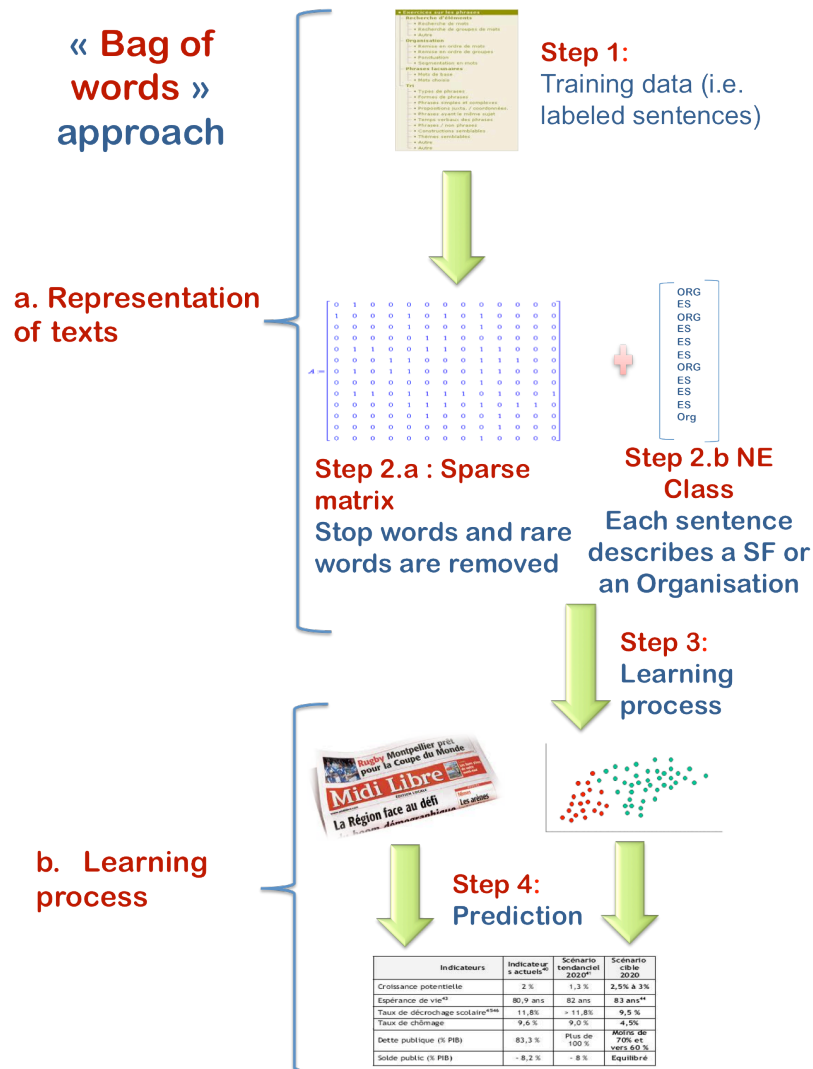


- **Extension:** How to analyse document with more precision?

**Example: Disambiguation** between **location** and **organisation** [Tahrat *et al.* WIMS'2013]



- **Disambiguation** between **location** and **organisation**



- **Disambiguation** between **location** and **organisation**

SVM			Naive Bayes		
	SF	OE		SF	OE
SF	103	35	SF	98	40
OE	44	90	OE	44	90
<i>Accuracy</i>		70.96%	<i>Accuracy</i>		69.12%

Features with ConceptOrg			Features with ConceptSpa			Both types of features		
	SF	OE		SF	OE		SF	OE
SF	108	30	SF	112	26	SF	113	25
OE	47	87	OE	19	115	OE	19	115
<i>Accuracy</i>		71.69%	<i>Accuracy</i>		83.45%	<i>Accuracy</i>		83.82%



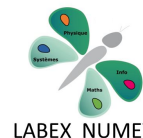


## Part 3

# Applications in agricultural domain

## Animal disease surveillance

**In collaboration with CMAEE lab**  
(Control of exotic and emerging animal diseases)





# Why the need of Epidemic Intelligence? (1/3)

More than **60% of the initial outbreak reports** come from unofficial informal and **heterogeneous sources**, including sources other than the electronic media, which **require verification** [Arsevska *et al.* ISVEE'2015]



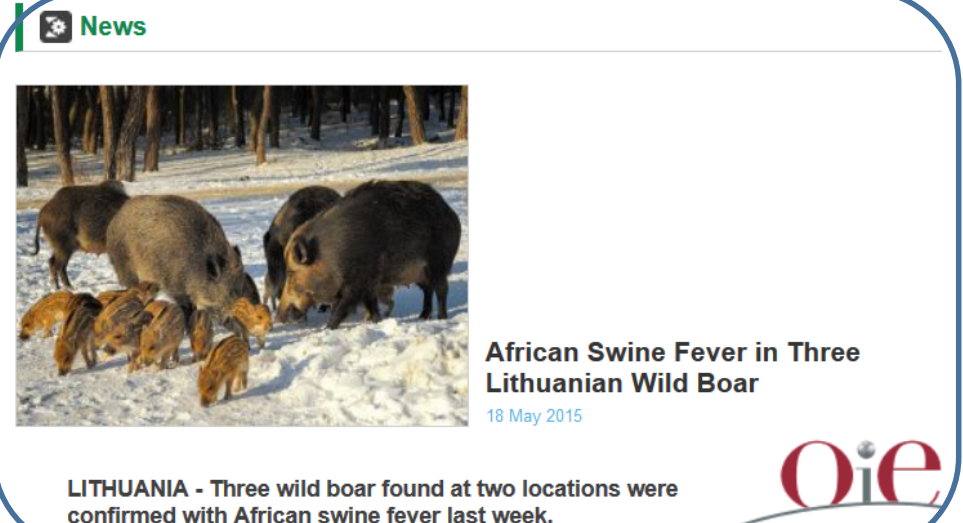
**INTERNATIONAL BUSINESS TIMES**  
MONDAY, JUNE 01, 2015 AS OF 2:24 PM CDT

Home Politics Economy Markets / Finance Companies Technology


TECHNOLOGY SCIENCE

**Unknown Disease Kills Kazakhstan's Rare Saiga Antelopes, Scientists Baffled**

By Kukil Bora [@KukilBora](#) on May 30 2015 7:21 AM EDT



**News**



**African Swine Fever in Three Lithuanian Wild Boar**  
18 May 2015

**LITHUANIA** - Three wild boar found at two locations were confirmed with African swine fever last week.

**Oie**

## Mysterious disease kills Nigerian patients within a day

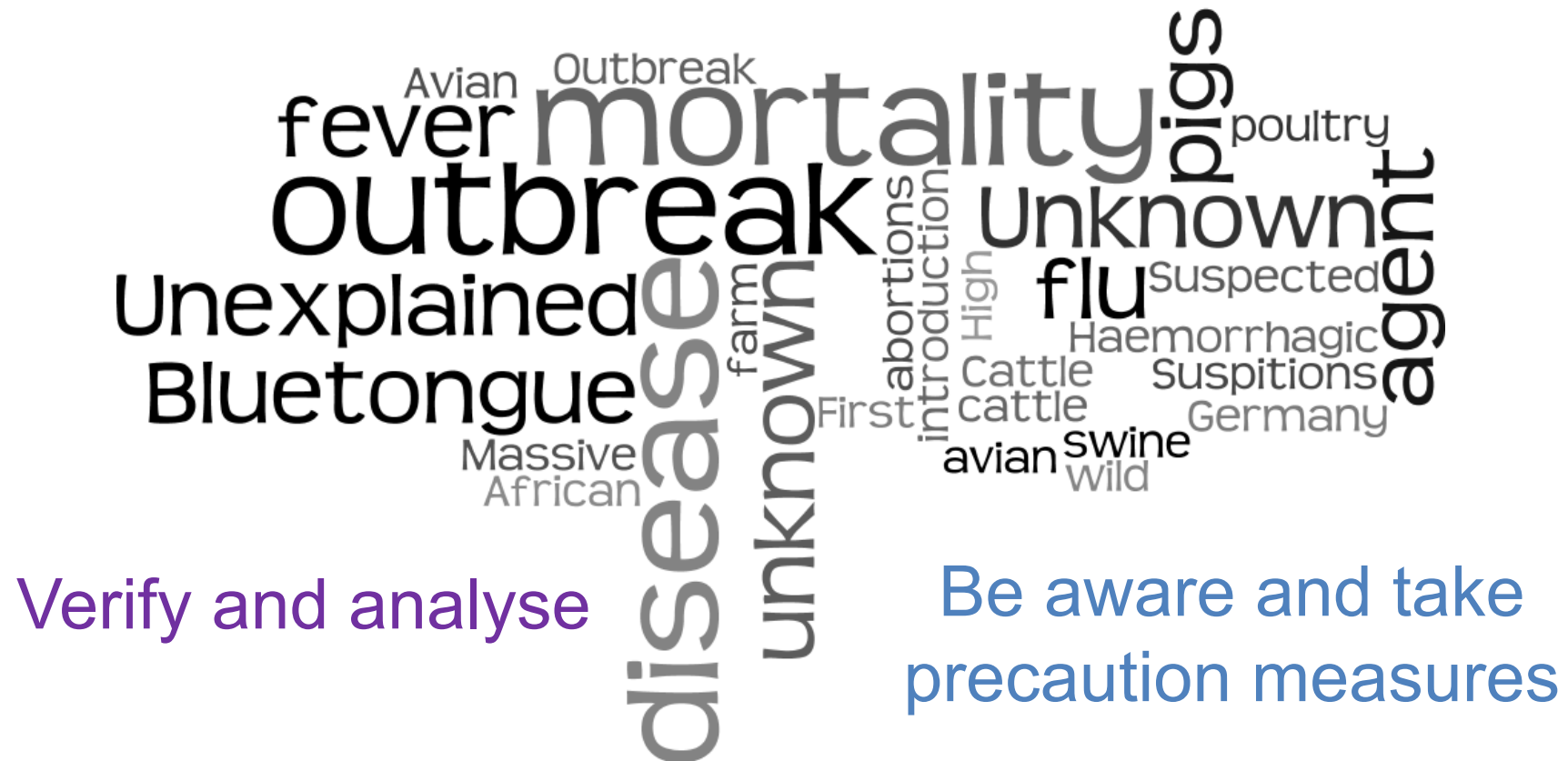
The unknown disease has so far killed 17 people in a southeastern Nigerian town and officials have ruled out Ebola.

18 Apr 2015 21:32 GMT | [Health](#), [Nigeria](#), [Africa](#)



# Why the need of Epidemic Intelligence? (3/3)

Identify signals of new and exotic animal diseases



# How to detect disease outbreak on the Web?

- **Four animal disease models:** African swine fever (ASF), Foot-and-mouth disease (FMD), Bluetongue (BTV), and Schmallenberg virus (SBV)
- **First model to study:** ASF



Industries | Wed Jul 23, 2014 9:54am EDT

Related: NON-CYCLICAL CONSUMER GOODS

## Poland investigates suspected case of African swine fever in farm pigs

WARSAW, JULY 23

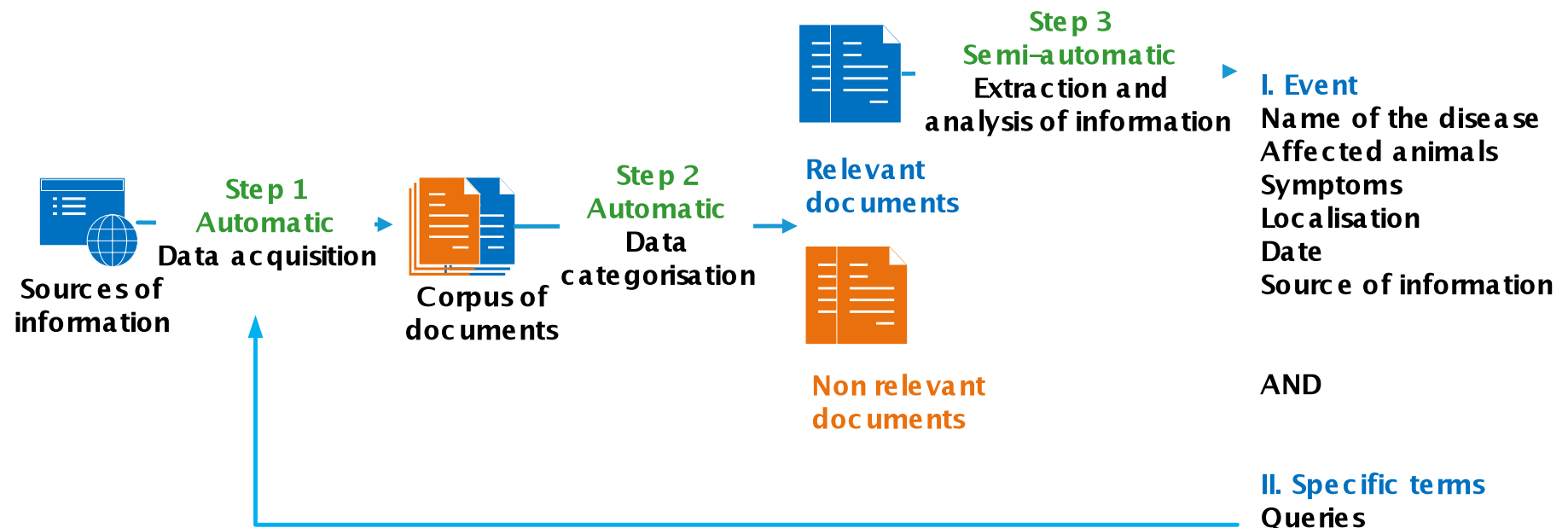
Polish local authorities said on Wednesday that preliminary tests have pointed to a case of African swine fever (ASF) among farm pigs in eastern Poland near the city of Bialystok.

The head of the Grodek county, Wieslaw Kulesza, told Reuters that preliminary results of tests showed that ASF was the cause of death of two-three farm pigs in the county.

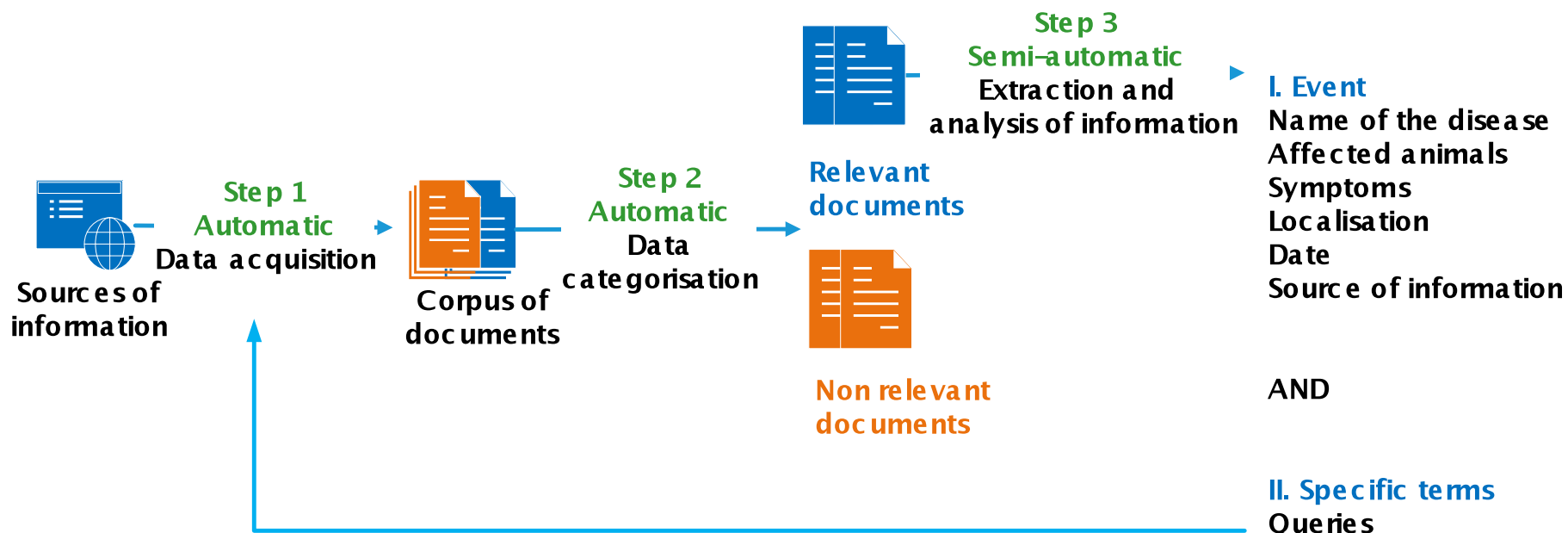
"We are marking the area," Kulesza said, adding that further steps, such as laying special mats, were being taken.

Poland's chief veterinary officer was unavailable for comment, while the county veterinary officer said a statement on the issue will be published later on Wednesday. (Reporting by Anna Wlodarczyk-Semczuk; Writing by Marcin Goettig)





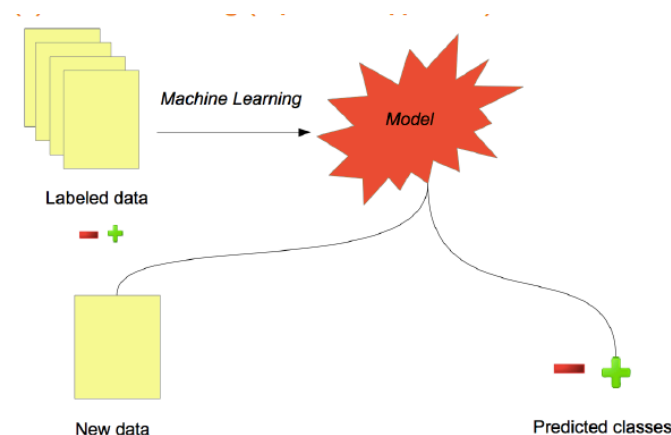
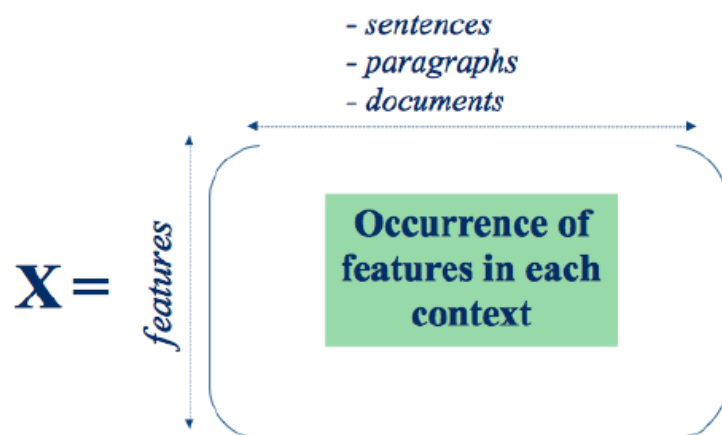
- Step 1: Data acquisition



<https://news.google.com/news/feeds?pz=1&cf=all&ned=en&q=Blue+tongue&output=rss>



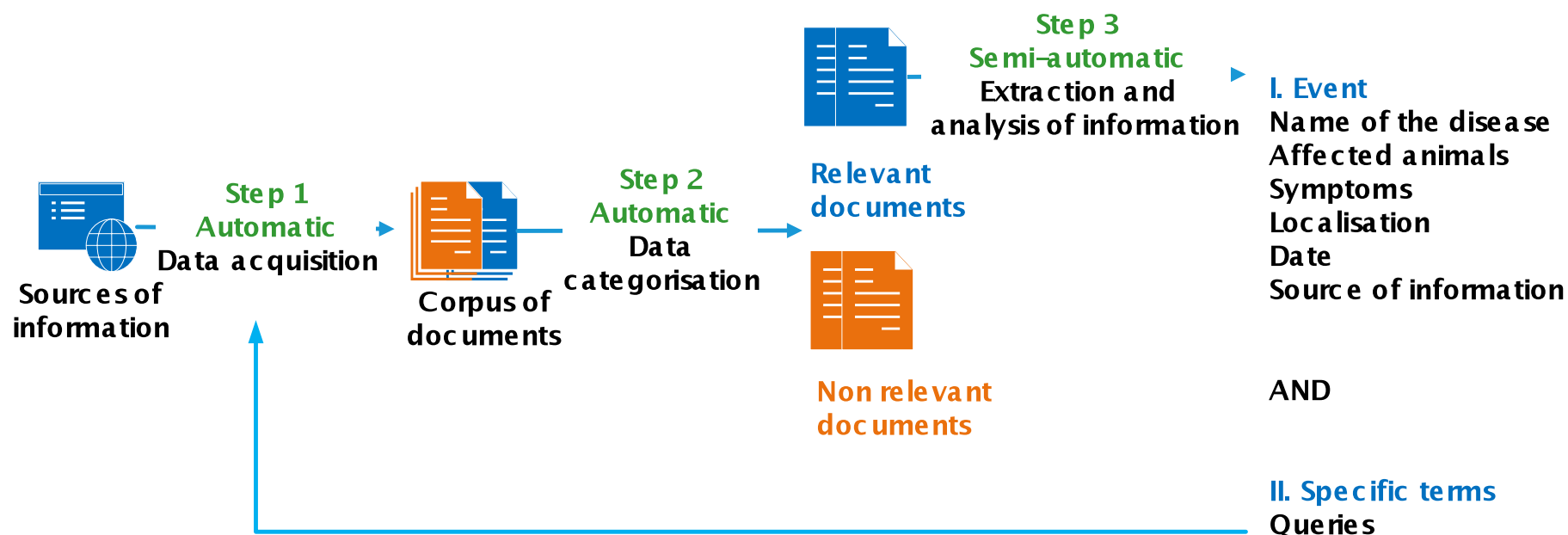
- Step 2: Data classification



Classification algorithm	Naïve Bayes			Support Vector Machine		
Performance	Recall	Precision	F-score	Recall	Precision	F-score
Class <i>disease</i>	0.724	0.766	0.744	0.657	0.68	0.669
<i>economy</i>	0.478	0.530	0.503	0.489	0.726	0.584
<i>general</i>	0.860	0.804	0.831	0.864	0.763	0.810
Weighted average	0.750	0.745	0.747	0.732	0.729	0.725



- Step 3: Information extraction and management



- **Step 3: Information extraction (I)**

**Aim:** Automatically detecting **key information** from **Web** news articles (country, species, diseases, number of cases, dates, ...)

“Since its initial appearance in **Poland** in **February 2014**, **72** cases of **African Swine Fever** have been detected in **wild boars** and there have been three outbreaks in **pigs**.” - <http://www.thenews.pl>

- Use **dictionaries** (Geonames, Heidelberg, disease names, species names, etc.), and data mining techniques in order to learn extraction rules.

**Rules associated with case numbers:**

(number)(species\_name,1-3) with support **26%** and confidence **83%**

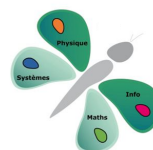
(number)(species\_name,1-2) with support **21%** and confidence **100%**





- **Step 3: Information extraction (I)**
  - **First results for the rule-based approach on the annotated corpus**
  - **Classification based on SVM** (features are rules)
  - **3 classes:** correct, incorrect, partial
  - 10-fold cross validation

Type	Accuracy (%)
Locations	70.6
Dates	71.2
Diseases	93.6
Cases	78.1
Species	89.5



LABEX NUMEV

**Julien Rabatel,  
LIRMM, Numev, France**



## • Step 3: Information extraction (I)

### Veille Sanitaire Internationale

Projet VSI Accueil Consultation Paramétrage

Annotation

Vous êtes connecté en tant que Elena Déconnexion

Jeux de données

annotations

532 article(s) Article précédent Article suivant

1. Rivers gov. eliminates chickens infected by flu 1 / 532

Date: 2015-01-20

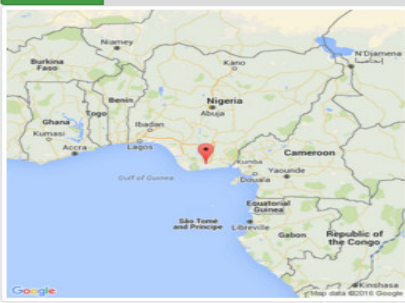
✓ Tout a été annoté Nouveau cas Bilan Non-pertinent

The 🇳🇮 Rivers State Government said on 📅 Tuesday that it had killed # hundreds of 🐔 fowls infected by the 🦠 Avian Flu in a privately owned farm in 📍 Port 🇳🇮 Harcourt .  
 The Commissioner for Agriculture , Emma 🇳🇮 Chinda , said that the farm had been quarantined and decontaminated .  
 He also said no human infection had been recorded .  
 " On 📅 January 🇳🇮 14 , we got a report from a farm that was worrisome . The report we got suggested that the farm may have been infected by the 🦠 highly pathogenic 🦠 avian influenza .  
 According to the commissioner , samples of the flu were taken to the Veterinary Research Institute in Vom , 🇳🇮 Plateau State .  
 " The result came out on 📅 January 🇳🇮 17 and it read positive of 🦠 highly pathogenic 🦠 avian influenza .  
 " On the basis of that , we had to take necessary steps . Apart from quarantining the farm , we had to depopulate the 🐔 birds in the farm to stop further spread " .  
 " Thereafter , we decontaminated the farm . We are containing the situation because officials of government and experts are on ground monitoring the situation " , he added .  
 Mr. 🇳🇮 Chinda said there was no need to panic because government was well equipped to handle the situation .  
 He said before the outbreak , they received information from the 🇳🇮 Federal Ministry of Agriculture on 🦠 avian influenza in 🇳🇮 Kano and a 🐔 bird market in 🇳🇮 Lagos .  
 " We were very much on alert and when it happened here , we handled the situation " , he said .  
 ( NAN )

29 candidat(s)  
 < Candidat précédent Candidat suivant >  
 6 / 29

LOCATION  
 " Port Harcourt "

✓ Correct 🇳🇮 Partiel ✗ Incorrect Non annoté

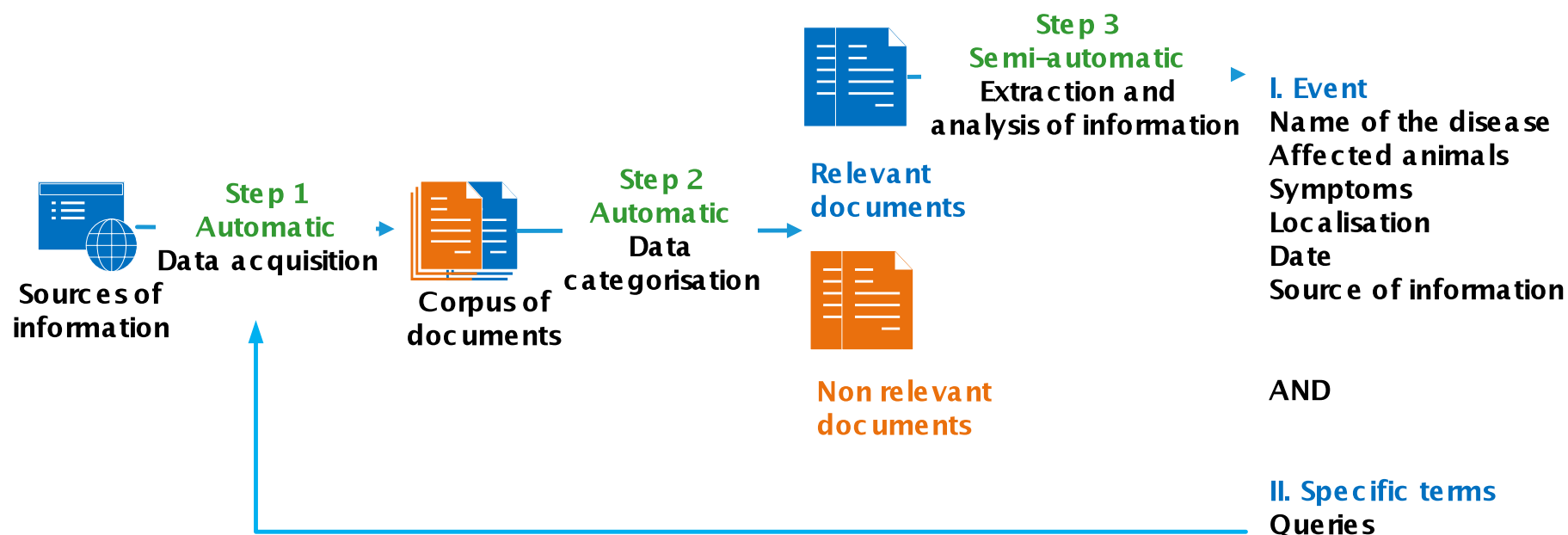


< Article précédent Article suivant >

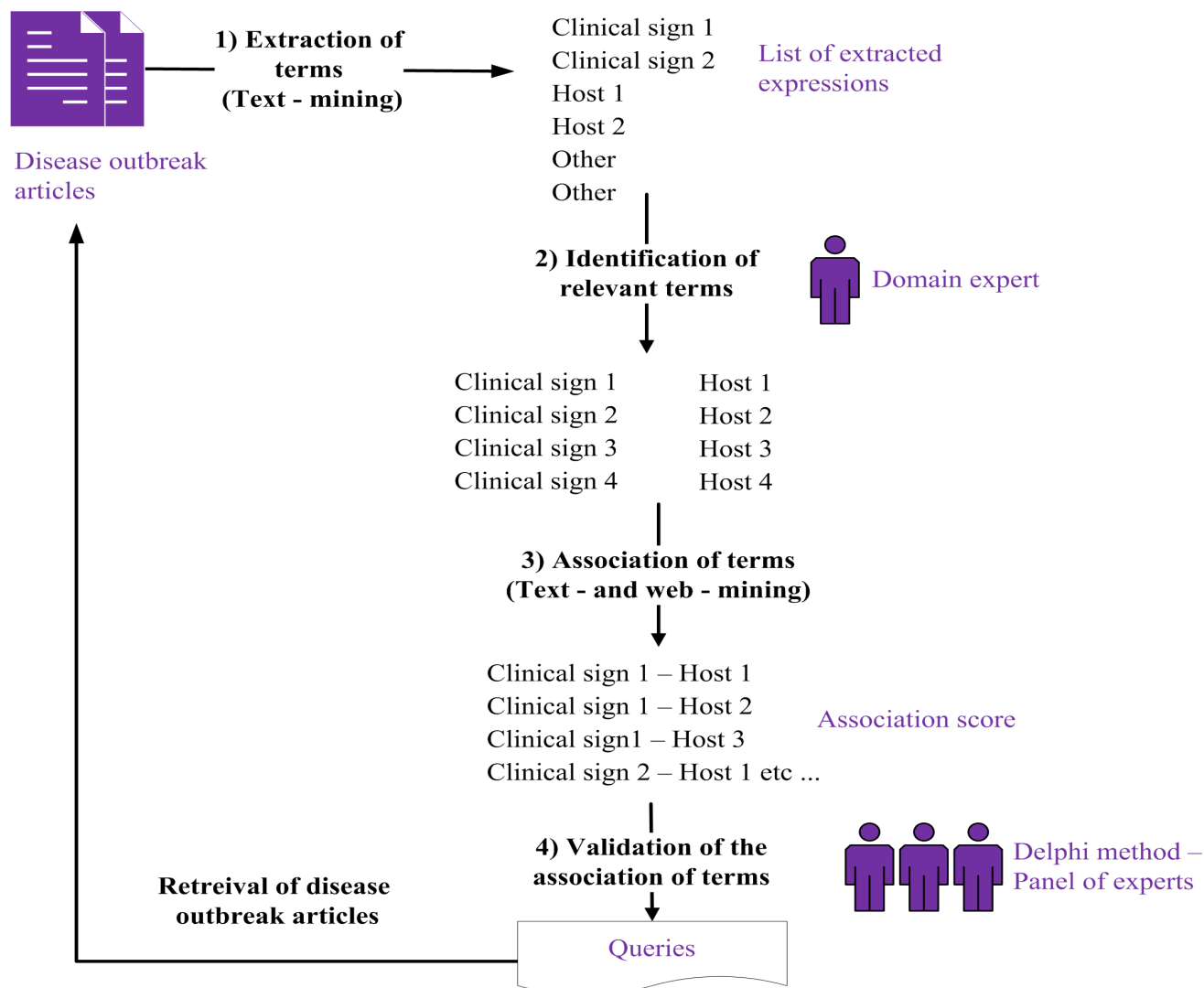
Max Devaud-Thomas Filol Copyright 2015 All Right Reserved.



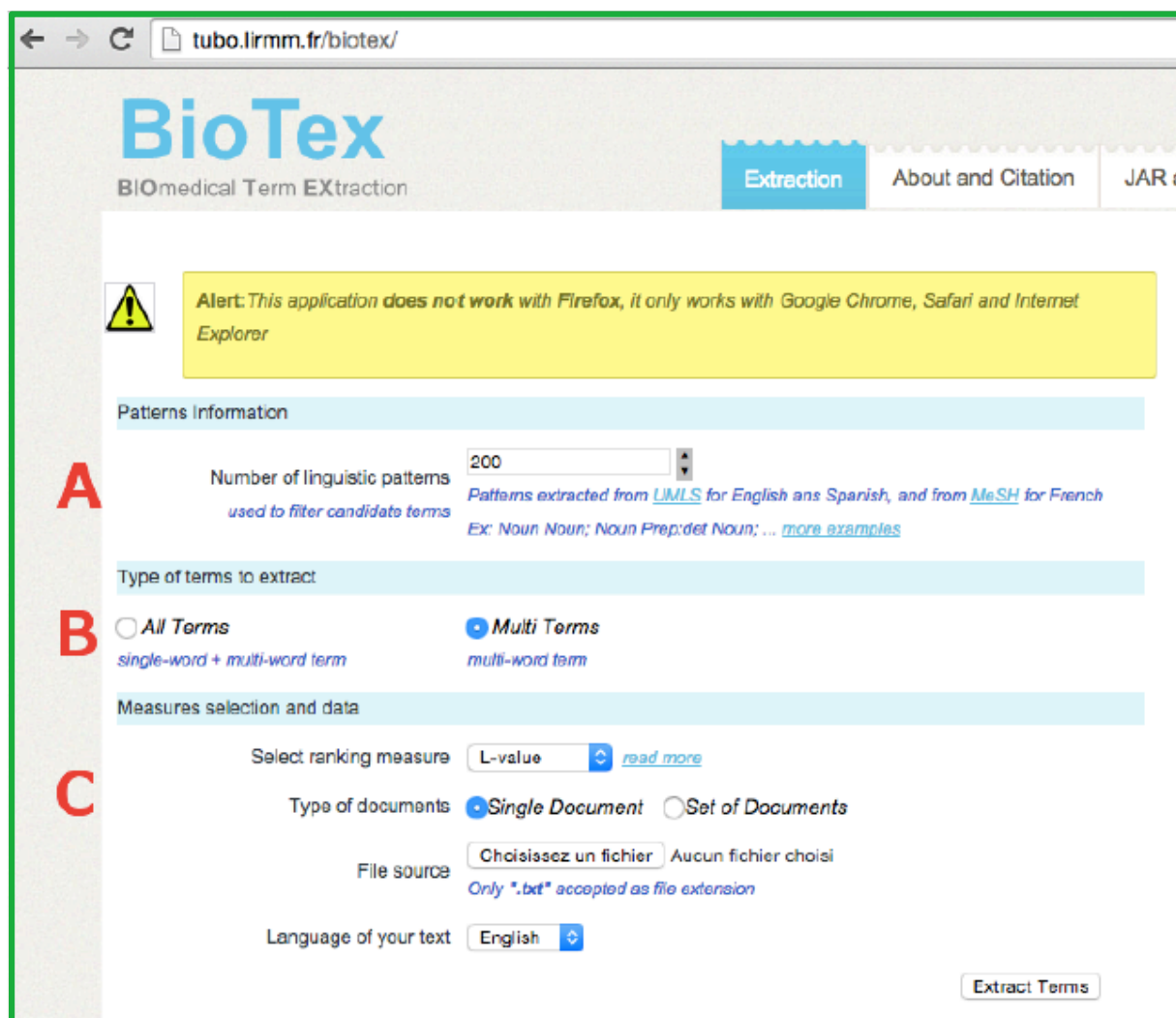
- Step 3: Information management (II)



## • II. Querying the Web



- II. Querying the Web: (a) *Terminology extraction*



The screenshot shows the BioTex web application interface. The browser address bar displays 'tubo.lirmm.fr/biotex/'. The page title is 'BioTex' with the subtitle 'BIOmedical Term EXtraction'. There are three tabs: 'Extraction' (active), 'About and Citation', and 'JAR a'. A yellow alert box states: 'Alert: This application does not work with Firefox, it only works with Google Chrome, Safari and Internet Explorer'. Below this is the 'Patterns Information' section, which includes a dropdown menu for 'Number of linguistic patterns' set to '200'. A note mentions patterns extracted from UMLS for English and Spanish, and from MeSH for French, with an example: 'Ex: Noun Noun; Noun Prep:det Noun; ... more examples'. The 'Type of terms to extract' section has two radio buttons: 'All Terms' (single-word + multi-word term) and 'Multi Terms' (multi-word term), with 'Multi Terms' selected. The 'Measures selection and data' section includes a 'Select ranking measure' dropdown set to 'L-value', a 'Type of documents' section with 'Single Document' selected, a 'File source' section with a file selection button and the note 'Only \*.txt\* accepted as file extension', and a 'Language of your text' dropdown set to 'English'. An 'Extract Terms' button is at the bottom right. Red letters A, B, and C are overlaid on the left side of the form, corresponding to the 'Patterns Information', 'Type of terms to extract', and 'Measures selection and data' sections respectively.



- II. Querying the Web: (b) *Terminology ranking*

## Statistics

- Frequency (TF) → **important** word

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Inverse Document Frequency (IDF) → **discriminant** word according the distribution in the corpus

$$IDF_i = \log \frac{|D|}{|d_j : t_i \in d_j|}$$

- Global value:

$$TF-IDF_{i,j} = TF_{i,j} \times IDF_i$$



- II. Querying the Web: (b) **Terminology ranking**

- BioTex Ranking [Lossio Ventura *et al.* IRJ'2016]:

$$LIDF-value(t) = P(t_{dom}) \times IDF(t) \times C-value(t)$$

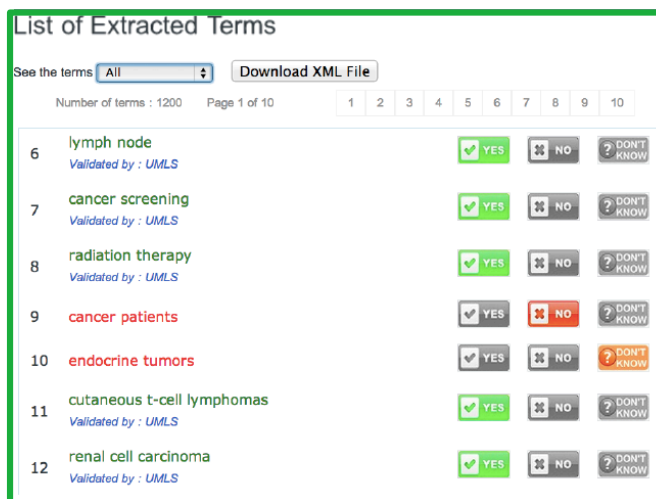
- A new ranking function to take into account the **heterogeneity** of the sources ( $S_i$ ) [Arsevska *et al.* CEA'2016]:

$$w(t) = \sum \alpha_i \times \frac{1}{rank_{S_i}(t)}$$

$$\text{with } \alpha_i \in [0,1] \text{ and } \sum \alpha_i = 1$$



- II. Querying the Web: (c) *Terminology validation*



The screenshot shows a web interface titled "List of Extracted Terms". It includes a dropdown menu set to "All" and a "Download XML File" button. Below this, it indicates "Number of terms : 1200" and "Page 1 of 10". A pagination bar shows numbers 1 through 10. The main content is a table of terms, each with a validation status and three buttons: "YES", "NO", and "DON'T KNOW".

Term	Validation Status	YES	NO	DON'T KNOW
6 lymph node <i>Validated by : UMLS</i>	✓ YES	YES	NO	DON'T KNOW
7 cancer screening <i>Validated by : UMLS</i>	✓ YES	YES	NO	DON'T KNOW
8 radiation therapy <i>Validated by : UMLS</i>	✓ YES	YES	NO	DON'T KNOW
9 cancer patients	✓ YES	YES	NO	DON'T KNOW
10 endocrine tumors	✓ YES	YES	NO	DON'T KNOW
11 cutaneous t-cell lymphomas <i>Validated by : UMLS</i>	✓ YES	YES	NO	DON'T KNOW
12 renal cell carcinoma <i>Validated by : UMLS</i>	✓ YES	YES	NO	DON'T KNOW

Using of a **Delphi method** [Arsevska *et al.* LREC'2016].

*Delphi method is to reach group consensus with experts (5 to 7 experts for each disease) when knowledge is not sufficient for a given scientific question.*





- II. Querying the Web: (c) **Terminology validation**

List of **extracted terms** identified to characterize Bluetongue virus (BTV) emergence.

Clinical signs	Term
General	<b>livestock deaths, general clinical signs, onset of weakness, excess mortality, fever outbreak</b>
Reproductive	<b>embryonic death, reproductive disorders, occurrence of abortion</b>
Hosts	Term
	<b>red deer, adult sheep, cattle herds, roe deer, cattle population, newborn calves, new born dairy calves, dairy calves, dairy ewes, pregnant ewes, cattle and goats, small ruminants</b>

*In bold are the terms proposed to experts for evaluation*



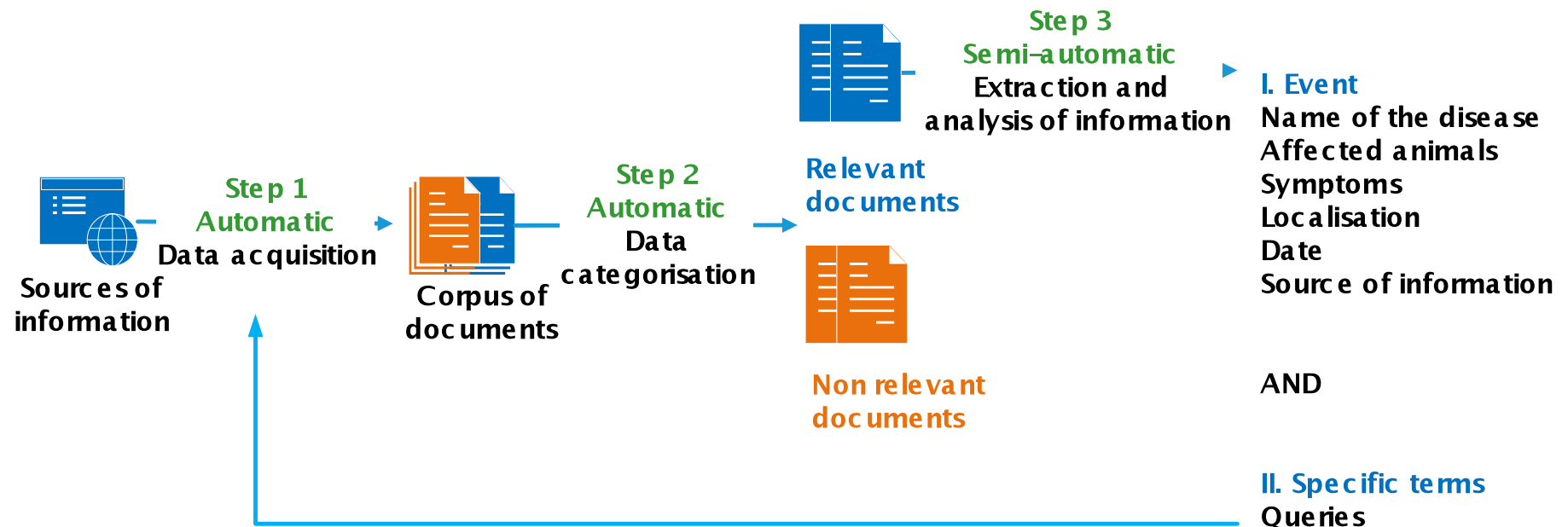
- II. Querying the Web: (d) **Association of terms**

$$D_{web}^{AND} = \frac{2 \times \text{hit}(h \text{ AND } cs)}{\text{hit}(h) + \text{hit}(cs)}$$

[Roche and Prince Informatica'2010 ; Arsevska *et al.* IJAEIS'2016]

Rank	Bluetongue <i>hosts / clinical signs</i>	Schmallenberg virus infection <i>hosts/ clinical signs</i>
1	general clinical signs / pregnant ewes	stillborn bovine foetuses / camels
2	livestock deaths / sheep	stillborn bovine foetuses / bison
3	embryonic death / cow	aborted foetuses / sheep
4	general clinical signs / sheep	deformed offspring / sheep
5	livestock deaths / cow	stillborn bovine foetuses / deer
6	livestock deaths / deer	aborted foetuses / cattle
7	fever outbreak / sheep	deformed offspring / cattle
8	embryonic death / sheep	stillborn bovine foetuses / calves
9	fever outbreak / cow	deformed offspring / lambs
10	embryonic death / pregnant ewes	acute bronchopneumonia / bison







## Part 3

# Applications in agricultural domain

## Sentiment analysis



## Methods in order to identify sentiments: *Towards a sentiment lexicon*

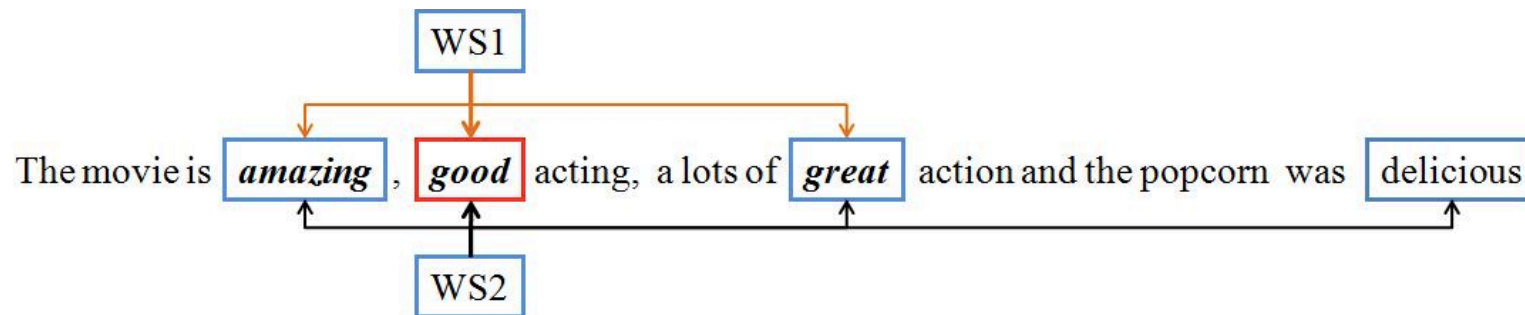
**Step 1:** choice of seeds related to opinions

$P = \{good; nice; excellent; positive; fortunate; correct; superior\}$

$N = \{bad; nasty; poor; negative; unfortunate; wrong; inferior\}$

Construction of 14 corpora related to a specific domain

**Step 2:** PoS, Association rules, choice of a window



## Step 3: Statistic selection and web mining.

Statistic measures that consist of measuring the association between **seed adjectives** and **candidate adjectives** association based on "hits" from the web (i.e. search engine) and contextual information

### Examples of learnt adjectives:

*great, hilarious, funny, happy, perfect, important, beautiful, amazing, complete, major, helpful*

### → Agriculture domain:

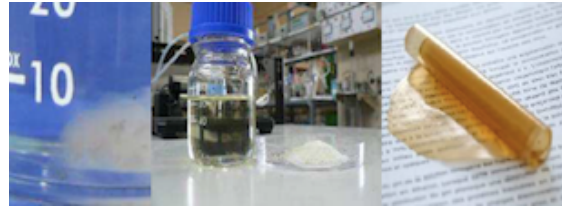
*{gmo; agricultural biotechnology; biotechnology for agriculture}*



Examples of learnt adjectives: *green, healthy, enthusiastic, creative, etc.*



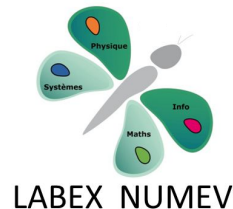
**Laura Vanessa Cruz, San Agustín University, Peru**



## Part 3

# Applications in agricultural domain

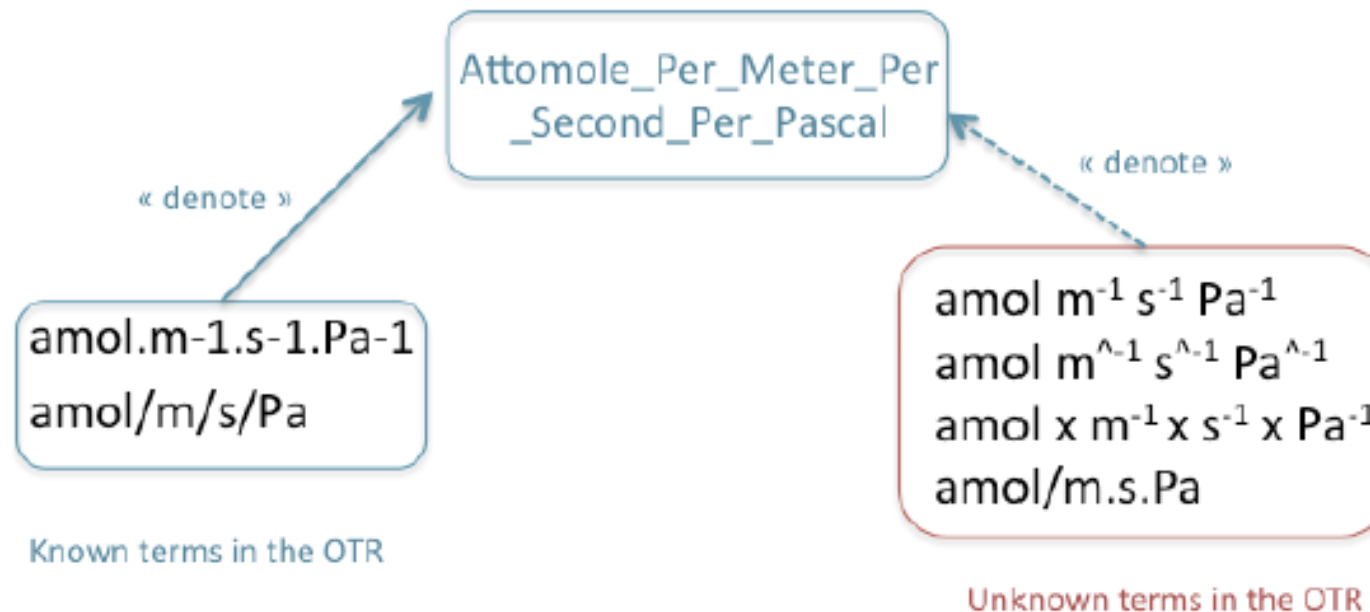
## Information Extraction from experimental data



Aim: **Knowledge management** in food science domain

Challenging issue: **Unit recognition and extraction**

[Berrahou *et al.* KDIR'2013 ; Berrahou *et al.* RNTI'2016]





## Method:

- **Locating unit** (*machine learning*)
- **Extracting unit** (*lexical similarity*)

$$SM_{DL}(u1, u2) = \max\left[0; \frac{\min(|u1|, |u2|) - DL(u1, u2)}{\min(|u1|, |u2|)}\right]$$

$$\in [0; 1]$$

Variant term	Reference	SMDc	SMDb
10e10 (cm3.m-1.sec-1.Pa-1)	10e10.cm3.m-1.sec-1.Pa-1	0.87	1
10e-14(cm3/m.s.Pa)	10e-14.cm3/(m.s.Pa)	0.89	1
10e-16cm3.cm/cm.cm2.s.Pa	(10e-16cm3.cm)/(cm2.s.Pa)	0.76	0.8
10e18 (mol.m/Pa.sec.m2)	10e18.mol.m/(Pa.sec.m2)	0.87	1
amol.m-1.s-1.Pa-1	amol.s-1.m-1.Pa-1	0.88	0.75
amol/m.s.Pa	amol/(m.s.Pa)	0.84	1
amol/m.sec.Pa	amol/(m.s.Pa)	0.69	0.75
cm3.um/m2.d.kPa	cm3.μm/(m2.d.kPa)	0.77	0.8



## Part 4

# Conclusions and future work



## New challenges of *Big Data*:

- **Matching different types** of documents (image/text, video/text, and so forth)



- **Integration of visual analytics** skills [Fadloun, Inforsid'2016]

